

1. Sequence Analysis

1.1 Plant Material

The *Triticum aestivum* landrace “Chinese Spring” was selected for sequencing as it is widely used for cytogenetic analysis¹ and physical mapping². A single seed descent line maintained for 8 generations and with an established provenance to original Sears lines, termed “CS42”, was provided by James Simmons, John Innes Centre. DNA and cDNA derived from RNA of this line was used for sequencing. *Triticum monococcum* accession 4342-96 was selected for sequencing, as it is a widely used community standard line for TILLING, physical mapping and genetic analysis. *Aegilops tauschii* ssp *stragulata* accession AL8/78 has been sequenced using 454 and SOLiD technology (Luo *et al.*, 2012, submitted). *Triticum aestivum* genomic DNA from the U.K. commercial varieties Avalon, Rialto and Savannah was also sequenced on the SOLiD platform and used in this project.

1.2 *Triticum aestivum* DNA and RNA isolation and cDNA synthesis

Genomic DNA was isolated from purified nuclei based on modification of an existing protocol³. A modified sucrose buffer SEB (Sucrose-based Extraction Buffer: 10% v/v TKE, 500 mM sucrose, 4 mM spermidine, 1 mM spermine tetrahydrochloride, 1.2g/L PEG 8000, 0.13% w/v sodium diethyldithiocarbamate and 0.2% v/v β -mercaptoethanol) was used instead of the recommended MEB buffer, which gave poor quality gDNA in wheat. 60-80g frozen leaf material was ground in liquid nitrogen and added to 1 litre SEB, and the protocol followed. Typically 200-500ug of high quality gDNA was extracted using this method. Total RNA from several tissues was extracted using either the RNeasy mini prep kit (Qiagen), or tri- reagent (Sigma) following the protocol online at www.mrcgene.com/tri/htm. RNA was extracted from seeds using protocol 2 from⁴. Between 0.5g and 1g of frozen material ground in liquid nitrogen with a cold mortar and pestle and extracted. RNA was treated with DNase (Roche), followed by 0.8ug/ul protease K (Roche, 20ug/ul stock), extracted by phenol/chloroform and precipitated with addition 1/10th v/v 3M sodium acetate pH 5.2, 1/1000th v/v glycogen (Roche, 20ug/ul stock) and 3 volumes of ethanol. RNA was subsequently purified using RNeasy MINElute clean up kit (Qiagen) and analysed with an Agilent Bioanalyser. mRNA was isolated using an Oligotex mRNA mini kit (Qiagen), using up to 250ug total RNA per column. cDNA was synthesizing from either 0.5ug mRNA or 3ug total RNA pools. First strand cDNA synthesis was performed using the reagents from the cDNA MINT-Universal kit (Evrogen), but using custom primers that carry a 1 base modification creating an Mmel site.

Mmel 3' Primer:

5' AAGCAGTGGTATCCAACGCAGAGTACTTTTTTTTTTTTTTTTTT 3'

Mmel PlugOligo adaptor:

5' AAGCAGTGGTATCCAACGCAGAGTACGGGGG 3' (Plus 3' Phosphorylation)

Two microliters of 1st strand cDNA was amplified by PCR using Mmel M1 primer: 5'AAGCAGTGGTATCCAACGCAGAGT 3'. Thermal cycler parameters were 95° 1 min followed by 10-15 cycles of 95° for 15 secs, 63° for 30 secs 72° for 3 mins, ending with 1 cycle of 72° for 10 mins. PCR was performed using the Advantage 2 PCR kit (Clontech). After amplification cDNA was purified on a Qiagen Minielute column and eluted in 30 ul EB. Purified cDNA (120ng) was used in the normalization procedure as described in the Evrogen TRIMMER kit. DSN enzyme⁵ was tested using the controls in the kit and usually required a dilution of between 2 and 4 fold for the desired level of digestion. Before DSN treatment of the cDNA sample it was denatured and annealed for 7 hours. The digestion was carried out for 25 minutes and stopped using the reagents in the kit. The material was purified using Qiagen Minelute clean up kit and eluted in 15ul. 3ul of this was used in a 50ul PCR reaction as before using Mmel M1 primer for 15 cycles. Following amplification the normalised product were purified using the QiaQuick PCR cleanup kit (Qiagen) and eluted in 50ul. It was then digested with 2 units of Mmel (New England biolabs) for 2 hours in a total volume of 100ul. The digested cDNA was re-purified twice using a Minelute clean up column and eluted in 25ul. The samples were then processed for 454 sequencing.

1.3 454 Pyrosequencing of *T. aestivum* DNA and cDNA

Random shotgun libraries were generated by nebulization of 5µg each of CS42 nuclear DNA and four cDNA pools and size selected using the double Spri option to obtain a fragment length distribution between 500-800bp. Adaptors were blunt-ligated to fragment ends and the final single-stranded DNA library was isolated via streptavidin bead binding to biotinylated adaptors followed by alkaline treatment. The library was quantitated by fluorometry using Quant-iT Ribo Green reagent (Invitrogen, Carlsbad, CA) prior to emulsion PCR amplification. A 1k-beta shotgun genomic library was prepared following the Rapid Library Preparation kit and following the manufacture's recommendations in the Manual GS FLX Titanium Series 1 K Beta (454 Life Sciences, Branford, CT). Two µg CS42 nuclear DNA was nebulized and size-selected by agarose gels to obtain a fragment size distribution between 800-1000bp to take advantage of the extra-long read length capabilities of the GS FLX 1 k-beta chemistry. Adaptors were blunt-ligated to fragments end and the final double stranded DNA library was quantitated via fluorometry using Quanti-iT Pico Green reagents (Invitrogen, Carlsbad, CA) prior to emulsion PCR amplification.

Both genomic shotgun libraries were clonally amplified via emulsion PCR by adding 0.5 molecule/bead per cup of emulsion, following manufacturer's recommendations employing

the GS FLX Titanium LV emPCR Kit (454 Life Sciences, Branford, CT). Following amplification, emPCR reactions were collected, and emulsions broken according to the manufacturer's protocols. Beads containing sufficient copies of clonally amplified library fragments were selected via the specified enrichment procedure and counted with a Z2 Coulter Counter (Beckman Coulter, Fullerton, CA) prior to sequencing.

Following emulsion PCR enrichment, beads produced using the titanium library were deposited into 2-region gasket format wells of a Titanium Series PicoTiterPlate device and 454 Sequencing was performed using the GS FLX Titanium Sequencing Kit XLR70 on the GS FLX instrument according to the manufacturer's recommendations (454 Life Sciences, Branford, CT). The enriched beads produced using the 1 k beta library were deposited on a the same format gasket but was used the GS FLX 1 k beta Sequencing Kit and a different sequencing script to carry out 350 cycles on an upgraded version of the GS FLX instrument. Image analysis, signal processing and base calling were performed using supplied system software. Standard Flowgram Format (sff) files output from base calling were employed in subsequent analysis.

1.4 SOLiD Sequencing of *T. aestivum* cv. Chinese spring, Avalon, Rialto and Savannah

A 2x50bp mate-paired library was constructed using CS42 according to the Applied Biosystems SOLiD™ 3 System Library Preparation Guide. Briefly, 5 µg of genomic DNA was fragmented by a HydroShear (Digilab Genomic Solutions Inc) to 2.0–3.5kb. The fragmented DNA was run on a 1% Agarose gel and DNA fragments between 2-3 kb were purified and end-repaired. MP adaptors were ligated to the sheared end-repaired DNA, which were then circularized by intra-molecular hybridization to protect the 3' overhangs of the MP Adaptors from self-annealing. Plasmid-Safe DNase was used to eliminate un-circularized DNA, yielding 1.3 µg of circularized DNA that was purified using Agencourt Ampure XP beads. Nick translation using *E. coli* DNA polymerase was performed for 10 minutes at 5°C to produce mate-paired tags of 100 bp. The nick translated DNA was digested with T7 exonuclease and S1 nuclease to cleave the mate-paired tags from the circularized template. P1-T and P2-T adaptors were ligated to the fragments to generate the 250-350 bp library. Finally, 250–300bp fragments were selected using E-Gel® system (Invitrogen) to ensure an average target length of around 100bp in the completed mate-paired library. This was then amplified using Library Primers 1 and 2 with Platinum Amplification Mix and 9 cycles of amplification.

Templated beads were prepared from the mate-paired library according to manufacturer's instructions using the ePCR kit v.3 and the Bead Enrichment Kit from Applied Biosystems (Life Technologies, Inc.) for SOLiD3 and the SOLiD™ EZ Bead™ System. The emulsion

was prepared from 0.5 pmol wheat CS42 mate-paired library added to the Aqueous Master Mix and P1 Beads to generate the aqueous phases. This was emulsified using the SOLiD™ EZ Bead™ Emulsifier (Applied Biosystems SOLiD™ EZ Bead™ Emulsifier Getting Started Guide) according to the manufacturer's recommendations (Applied Biosystems SOLiD™ EZ Bead™ Amplifier Getting Started Guide). The beads were washed and enriched using the SOLiD™ EZ Bead™ Enricher following the manufacturer's recommendations (the Applied Biosystems SOLiD™ EZ Bead™ Enricher Getting Started Guide). After enrichment, terminal transferase was used to modify the 3'ends of the library. Library quality was checked using the Workflow Analysis kit from Applied Biosystems (Life Technologies, Inc.) Templated beads were deposited on slides according to manufacturers' instructions using the Bead Deposition kit from Applied Biosystems (Life Technologies, Inc.). Two full slides (1 full run) of the 2×50bp Chinese spring mate-paired library were sequenced according to the Applied Biosystems SOLiD 3 System Instrument Operation Guide and all other runs (9.5 runs, 19 slides) of Chinese spring, Avalon, Rialto and Savannah were sequenced on the SOLiD 4 platform. Library construction followed the Applied Biosystems SOLiD™ 5500 series Library Preparation Guide, as this improved on the SOLiD 4 protocol.

| Repeat | No. of reads | % of total |
|-------------------------|---------------------|-------------------|
| DNA transposons | 32,760,567 | 14.934 |
| Helitron | 531,038 | 0.242 |
| TIR | 32,094,850 | 14.630 |
| HAT | 91,884 | 0.042 |
| Harbinger | 748,318 | 0.341 |
| Mariner | 2,223,675 | 1.014 |
| CACTA | 28,035,574 | 12.780 |
| Mutator | 975,821 | 0.445 |
| Unknown | 19,578 | 0.009 |
| Unknown | 134,679 | 0.061 |
| Retrotransposons | 139,831,934 | 63.742 |
| SINE | 8,796 | 0.004 |
| LINE | 1,798,035 | 0.820 |
| LTR | 138,025,103 | 62.918 |
| Unknown | 3,269,006 | 1.490 |
| Gypsy | 96,598,453 | 44.034 |
| Copia | 38,157,644 | 17.394 |
| Unknown | 2,680,931 | 1.222 |

Supplementary Table 1. Repeat composition of 454 reads.

The 454 reads were BLASTN screened against the TREP repeat database and the number of reads matching repeats at e-5 or less were recorded.

1.5 Data sets and availability

| Analysis | Genome | Platform (& library) | Size of dataset | Accession |
|---|--|---|--|--|
| Orthologous group assembly (OA) | <i>T. aestivum</i> (Chinese spring) | 454 GS FLX Titanium and 454 GS FLX+ (Genomic fragment) | 85Gb | EBI Study: ERP000319, PRJEB217 |
| Low copy number assembly (LCG) | | | | Download and BLAST search 454 reads at www.cerealsdb.uk.net |
| BLASTN comparison with <i>B. distachyon</i> genes | | | | |
| Homeologous sequence assignment (using raw reads) Definition of training set | <i>T. aestivum</i> (Chinese spring) chromosomes 1A, 1B, 1D | Chromosome sorted 454 GS FLX Titanium (Genomic fragment repeat masked) | 1A: 287Mb 1B: 392Mb 1D: 375Mb | EBI Study: ERP000446 Wicker <i>et al.</i> , 2011 |
| Homeologous sequence assignment (using raw reads) | <i>Ae. tauschii</i> (representing D genome) | 454 GS FLX Titanium (Genomic fragment) | 12.8Gb | NCBI archive: SRA052214 |
| Homeologous sequence assignment (pre-assembled data) | <i>Ae. speltooides</i> (representing B genome) | De novo assembly of cDNA reads | 151Mb | Trick & Bancroft, unpublished |
| Homeologous sequence assignment (assembled reads) | <i>T. monococcum</i> (representing A genome) | Illumina GAIIx/HiSeq, (Genomic 100bp paired-end/36bp paired-end) | 3.7Gb | NCBI archive: SRP004490.3 |
| Homeologous SNP identification (using raw reads) | | | 401Gb | |
| Homeologous SNP identification (using raw reads) | <i>Ae. tauschii</i> (representing D genome) | SOLiD 4 (Genomic fragment) | - | Luo <i>et al.</i> , 2012, submitted |
| Homeologous SNP identification (using raw reads) | <i>T. aestivum</i> (Chinese spring with 4 more varieties, see Suppl. Online Material) | SOLiD 3 and SOLiD 4 (Genomic 50bp mate-pair) | 15.2 billion reads | EBI Study: ERP001493 |

| | | | | |
|---|-------------------------------------|---|-------|----------------------|
| Associating <i>T. aestivum</i> transcriptome with OG assembly (assembled reads) | <i>T. aestivum</i> (Chinese spring) | 454 GS FLX Titanium & GS FLX+ (cDNA fragment) | 1.6Gb | EBI Study: ERP001415 |
|---|-------------------------------------|---|-------|----------------------|

Supplementary Table 2. Summary of all datasets used, listed by analysis type.

2. Sequence Assembly

2.1 Low Copy Genome Assembly

Prior to assembly reads matching the Triticeae Repeat Database (TREP) were removed (megablast -e 1e-05, -F F), as were reads matching the wheat chloroplast (GenBank: NC_002762.1) and mitochondrial (GenBank: NC_007579.1) genomes (megablast -e 1e-15, -F F) and reads matching the ribosomal database SILVA⁶ (BLASTn -e 1e-05, -F F). These filters removed ~60% of the reads. Of the remaining ~40%, a subset of 17% was identified as genic by comparison to genes from *Brachypodium distachyon* (v1.2), our wheat transcriptome 454 reads and publicly available EST data from the DFCI Triticum aestivum Gene Index (TaGI, release 11.0) and the NCBI UniGene set (build 56) (all using megablast -e 1e-05, -F F). The 40% (~87 million) of reads remaining after filtering was assembled using gsAssembler from the Newbler package (development version 2.6pre) using the “-large” parameter.

| | Ortholome gene assembly (OA) (mi 99%) | Low-copy number genome assembly (LCG) |
|-----------------------------------|--|--|
| # of singletons | 1,222,242 (31%) | 7,683,490 (9%) |
| # of assembled reads | 2,689,502 (67%) | 70,635,013 (81%) |
| # reads excluded from assembly | 90,252 (2%) | 8,371,631 (10%) |
| # of repeat reads | 1,025 | 3,121,988 |
| # of outlier reads | 89,190 | 2,652,442 |
| # too short reads | 39 | 2,597,201 |
| # of assembled contigs | 172,039 | 5,321,847 |
| Total contigs + singletons | 1,394,281 | 13,005,337 |
| Total sequence (bp) | 630,756,335 | 5,422,739,356 |
| Min / max length (bp) | 52 / 7,312 | 40 / 21,721 |
| Mean length (bp) | 452.39 | 416.96 |
| N50 / N90 (bp) | 479 / 326 | 621 / 222 |

Supplementary Table 3. Newbler assembly statistics for the wheat Low Copy Genome (LCG) assembly and Ortholome- based Assembly (OA).

2.2 Construction of an orthologous grass gene set

To define gene family clusters from the sequenced genomes of three grasses from diverse grass sub-families and publicly- available barley full-length cDNAs, the OrthoMCL software^{7,8} version 1.4 was used. In a first step, pairwise sequence similarities between all input protein sequences were calculated using BLASTP with an e-value cut-off of 1e-05. Markov clustering of the resulting similarity matrix was used to define the ortholog cluster structure, using an inflation value (-I) of 1.5 (OrthoMCL default).

The input datasets were:

Brachypodium distachyon: v1.2 MIPS

Sorghum bicolor: v1.4 MIPS

Oryza sativa: RAP2

Hordeum vulgare: flCDNAs CD-HIT clustered

Splice variants were removed from the data set, keeping the longest protein sequence prediction, and data sets were filtered for internal stop codons and incompatible reading frames. A total of 86,944 coding sequences from these four grasses were clustered into 20,496 gene families. 9,843 clusters contained sequences from all four genomes. An overview of the cluster distribution is shown in Supplementary Figure 1. One

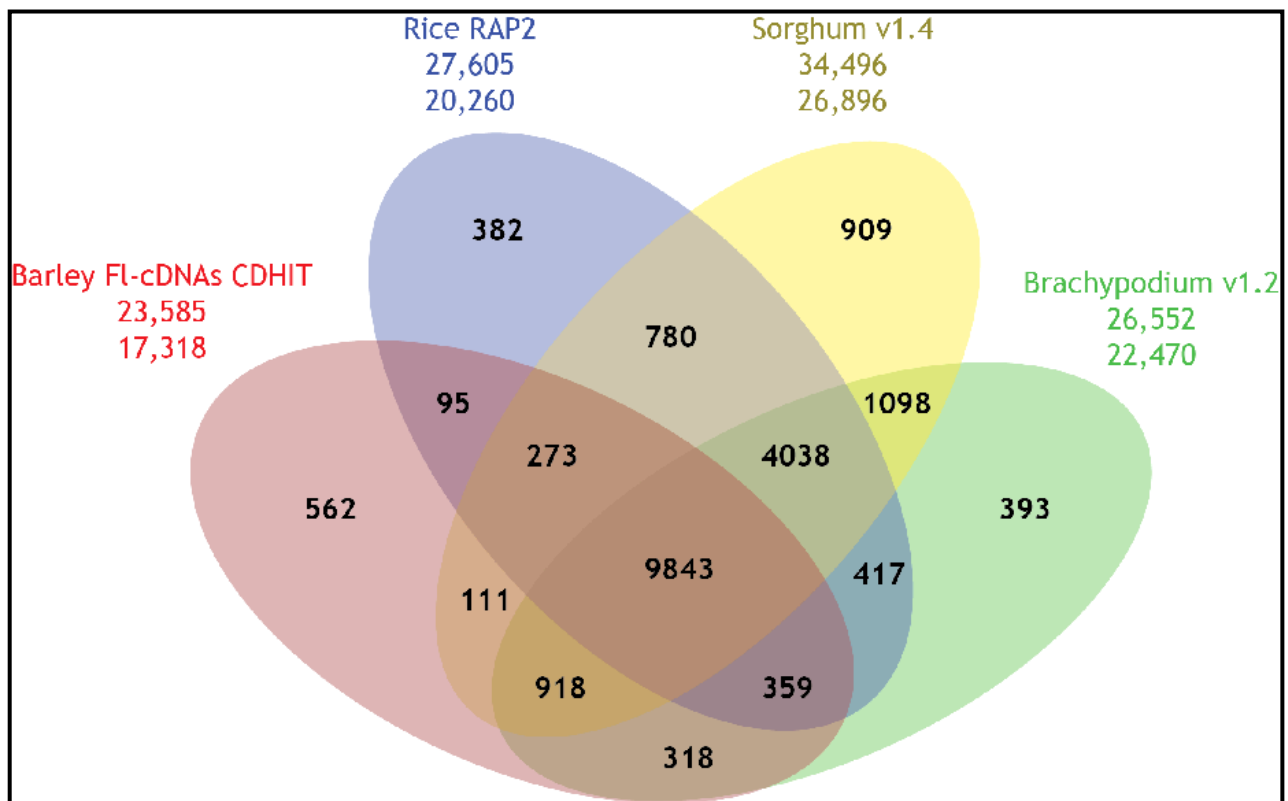
representative gene model was selected from the OrthoMCL analysis for each of the 20,496 orthologous gene clusters, using the following procedure:

1. BLASTX of all contigs from the CS42 LCG assembly against all grass genes used in the OrthoMCL analysis;
2. Identify the gene in each cluster that identifies the most DISTINCT wheat contigs;
3. If genes pool the SAME number of wheat contigs, select the one with the longest protein sequence as the representative gene model.

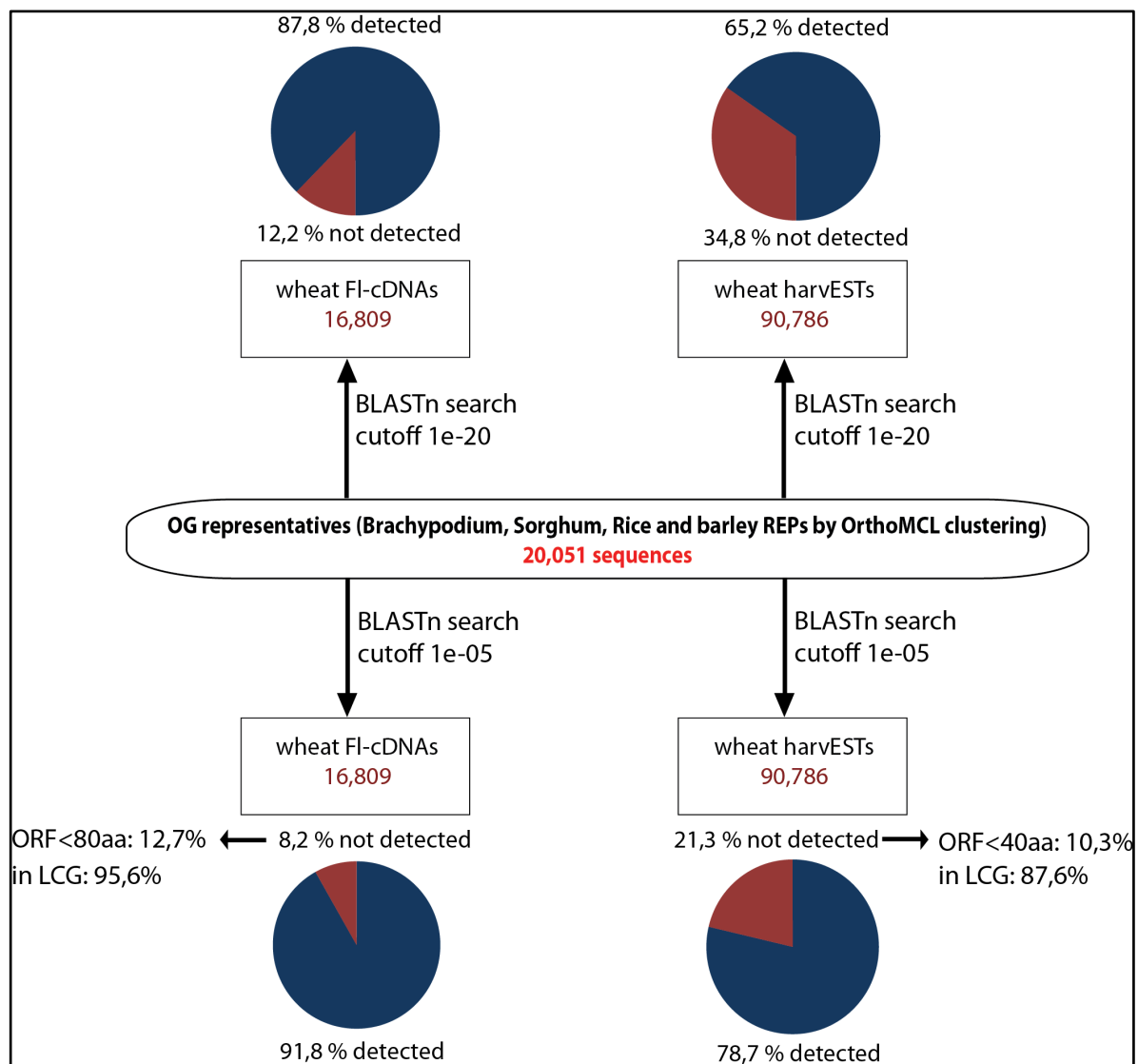
Genes with associated wheat contigs were identified for all but 445 gene clusters resulting in 20,051 representative gene models for all subsequent analyses.

| OG Representative | Number of Clusters |
|--------------------------|---------------------------|
| Brachypodium | 7,996 |
| Barley flcDNAs | 5,337 |
| Sorghum | 3,582 |
| Rice | 3,136 |
| without representative | 445 |
| Σ 20,496 | |

Supplementary Table 4. Definition of Ortologous Gene Representatives (OGs).



Supplementary Figure 1. The Grass Ortholome. The orthologous gene (OG) set. In each intersection of the Venn diagram the number of gene-groups ("families") are represented. The first number under the species name counts for the total number of genes used for a particular species, the second one gives the number of genes in clusters for that organism.



Supplementary Figure 2. The set of OG Representatives was compared to publicly available wheat FI-cDNAs⁹ and wheat ESTs (HarVEST, version 1.19, <http://harvest.ucr.edu/>) using BLASTn at two cutoffs. Wheat FI-cDNA and harvEST sequences not detected in the OG Representative set (using BLASTN evalue cutoff 1e-05; 8,2% resp. 21,3%) were searched against the LCG assembly to assess their representation. Between 95,6% (FI-cDNA) and 87,6% (harvESTs) of the sequences not matching an OG Representative were found in the LCG. The remaining fraction could be accounted for by contamination and non-genic transcription.

2.3 Ortholome Assembly

2.3.1 Preprocessing of 454 sequence data

The public raw whole-genome shotgun 454 reads were downloaded from <http://www.cerealsdb.uk.net> and repeat masked using Vmatch (<http://www.vmatch.de>) against the MIPS-REdat Poaceae v8.6.2 repeat library by applying a minimum identity

cutoff of 70%, minimum length 100bp, seed-length 14, exdrop 5 and e-value 0.001. Finally, reads were filtered for continuous stretches of at least 50bp un-masked sequence.

| Mapping of repeat masked 454 reads to Ortholome | |
|--|----------------------|
| Raw whole-genome sequencing data (454 reads) (source: http://www.cerealsdb.uk.net) | 213,098,052 |
| total sequence [bp] | 82,801,349,875 |
| minimum / maximum length [bp] | 18 / 2,032 |
| mean length [bp] | 388.56 |
| N50 / N90 [bp] | 461 / 284 |
| Repeat content [bp] | 62,290,705,717 (75%) |
| # of 454 reads remaining after repeat masking and filtering | 65,851,441 |
| total sequence [bp] | 23,500,630,080 (28%) |
| minimum / maximum length [bp] | 50 / 2,032 |
| mean length [bp] | 356.87 |
| N50 / N90 [bp] | 460 / 260 |
| Remaining repeat sequence [bp] | 3,273,503,565 (13%) |
| # of mapped reads | 4,058,985 (6%) |
| # of unique mapped reads | 2,740,044 (68%) |
| # of multiple mapped reads | 1,318,941(32%) |
| # of OG representatives matched by first-best hit of mapped reads (incl. TE-related OG representatives) | 19,483 (99%) |
| Extraction of quality information for mapped 454 reads (FASTQs from SRA) | |
| # with truncated quality associated sequences | 846,554 |
| # extended quality associated sequences | 17,746 |
| # without any quality information | 56,987 |
| # of reads with associated quality information used for sub-assemblies | 4,001,998 |
| # of OG representatives matched by mapped reads with quality information (incl. TE-related OG representatives) | 19,482 |

Supplementary Table 5. Preprocessing of 454 reads for ortholome-centric sub-assemblies.

2.3.2 Mapping of genomic sequence reads to OG representatives

Nucleotide sequences of repeat-masked and filtered 454 reads were compared against the protein sequence of OG Representatives using BLASTX. Returned hits were filtered requiring $\geq 80\%$ sequence identity for barley, $\geq 75\%$ for *Brachypodium* and $\geq 70\%$ for rice or

sorghum gene representatives spanning at least 30 amino acids. In case of matches to multiple OG Representatives only the first-best-hit of the sequence read was considered. Finally, genomic sequence reads were assigned to the matched OG Representative.

2.3.3 Gene-centric sub-assemblies

Unmasked sequences and corresponding quality information were extracted and grouped for all 454 reads that had been mapped on an OG Representative. For each group individual assemblies of the extracted sequences were computed using the Roche GS *de novo* Newbler assembler software v2.5.3. In this assembly we required a minimum overlap length (-ml) of 40bp and varied the minimum overlap identity (-mi) using 97%, 99% and 100%. Only contigs with a minimum length of 100bp were considered for further analysis (-l 100). The status of each read in the assembly was analyzed by consideration of the status file "454ReadStatus.txt" and reads that were marked as singletons were combined with contig sequences forming the set of sub-assembly sequences.

| | mi 97% | mi 99% | mi 100% |
|---|-----------------|-----------------|-----------------|
| Newbler sub-assembly statistics | | | |
| # of singletons | 887,615 (22%) | 1,222,242 (31%) | 1,696,740 (42%) |
| # of assembled reads | 3,038,943 (76%) | 2,689,502 (67%) | 2,057,928 (51%) |
| # reads excluded from the assembly | 75,440 (2%) | 90,254 (2%) | 247,330 (6%) |
| # of repeat reads ¹ | 899 | 1,025 | 480 |
| # of outlier reads ² | 74,502 | 89,190 | 246,811 |
| # too short reads ³ | 39 | 39 | 39 |
| # of assembled contigs | 205,817 | 172,039 | 120,501 |
| # of sub-assemblies used for copy number analysis (contigs + singletons) | 1,093,432 | 1,394,281 | 1,817,241 |
| total sequence [bp] | 497,965,174 | 630,756,335 | 793,978,129 |
| minimum / maximum length [bp] | 52 / 7,415 | 52 / 7,312 | 52 / 4,386 |
| mean length [bp] | 455.41 | 452.39 | 436.91 |
| N50 / N90 [bp] | 482 / 323 | 479 / 326 | 471 / 322 |
| Re-alignment of sub-assemblies to OG Representatives | | | |
| # of re-aligned sub-assemblies | 1,019,315 (93%) | 1,338,548 (96%) | 1,775,454 (98%) |
| # of re-aligned contigs | 153,619 | 143,193 | 109,802 |
| # of re-aligned singletons | 865,696 | 1,195,355 | 1,665,652 |
| # of OG Representatives with accepted, re-aligned sub-assemblies (incl. TE-related OG Representatives) | 19,429 | 19,467 | 19,475 |
| # of OG Representatives which are associated to TE and removed manually | 149 | 149 | 150 |
| ¹ the read was either inferred to be repetitive early in the assembly process (>70% of the read's seed hit to at least 70 other reads) ² the read was identified as problematic (e.g. chimeric sequences or assembler artifacts) ³ the read was too short to be used (<50bases and longer than the value of the minlen parameter used) | | | |

Supplementary Table 6. Newbler sub-assembly and re-alignment statistics.

2.4 Re-alignment of sub-assemblies to OG representatives

Sub-assembly sequences were aligned to the protein sequence of the OG Representative by BLASTX. Hits were filtered for $\geq 80\%$ sequence identity for barley, $\geq 75\%$ for *Brachypodium* and $\geq 70\%$ for rice or sorghum gene representatives spanning at least 30 amino acids. If multiple high scoring segment pairs (HSPs) were returned, only HSPs matching on the same strand such as the first-best-HSP were considered. To focus subsequent work on essentially complete wheat sub-assemblies, only sub-assemblies covering OG Representatives by at least 70% were taken forward for further analysis.

2.5 Prediction of wheat gene copy numbers

The wheat gene copy number was determined separately for each OG Representative. A term called the position-specific hit-count profile was determined by counting the number of mapped sub-assemblies located at a specific amino acid position of the template sequence. By only considering sequence positions of the OG Representatives that are tagged by one or more sub-assemblies, the percentage of sequence positions with evidence of x distinct sub-assemblies was determined, where x ranges from 1 to the maximum hit-count in the profile. From the distribution curve of x values, a coverage cut-off of C was defined as the minimum fraction of the covered OG Representatives that is covered by an entire sub-assembly or sequence-related sub-assemblies. Thus, the gene copy number is predicted as the maximum hit count assigned to $C=70\%$ of the OG Representative. OG Representatives with gene copy numbers of >75 were not considered further as these were generally associated with repeats.

2.6 Estimation of the gene retention rate in wheat and *Ae. tauschii*

The gene retention rate in wheat is a measure of the predicted wheat gene copy number relative to their gene family sizes in the sequenced diploid reference species, as determined by OrthoMCL analysis. We first determined the number of gene copies of the OG Representatives in their originating sequenced diploid genomes clustering genes related to the OG using OrthoMCL analysis. To calculate the gene retention rate, the wheat sub-assembly copy number of the OG was paired with the reference gene family size, and a polynomial fit of the data set was calculated using median locally-weighted polynomial regression (<http://stat.ethz.ch/R-manual/R-patched/library/stats/html/lowess.html>), incorporating OG Representatives with a maximum orthologous gene family size of ten. The steepness of the regression fit indicates gene retention over the whole sample size. We calculated the gradient at each data point of the polynomial approximation ($\Delta Y/\Delta X$), and finally estimated the gene retention rate as the mean gradient of the curve as well as its standard deviation.

| | mi 97% | mi 99% | mi 100% |
|---|---------------------|---------------------|---------------------|
| Reference used for determination of wheat gene copy number | | | |
| # of OG Representatives with $\geq 70\%$ coverage by re-aligned sub-assemblies (TE-cleaned) | 11,374 (59%) | 12,518 (64%) | 13,030 (67%) |
| # of sub-assemblies used for calculation of wheat gene copy number | 547,104 | 761,470 | 1,088,927 |
| total sequence [bp] | 258,749,616 | 353,579,457 | 479,847,207 |
| minimum / maximum length [bp] | 80 / 7,415 | 79 / 7,312 | 79 / 4,386 |
| mean length [bp] | 472.94 | 464.34 | 440.66 |
| N50 / N90 [bp] | 487 / 335 | 483 / 332 | 472 / 324 |
| Wheat gene copy number and gene retention rate | | | |
| # OG Representatives with wheat copy number ≤ 75 | 11,355 | 12,481 | 12,948 |
| Mean gene copy number | 3.62 ± 5.44 | 4.71 ± 6.33 | 6.42 ± 7.42 |
| Median gene copy number | 2.0 | 3.0 | 4.0 |
| Gene retention rate | 1 : 1.34 ± 0.23 | 1 : 1.83 ± 0.34 | 1 : 2.70 ± 0.29 |
| # OG Representatives with significantly expanded gene copy number | 512 (4%) | 560 (4%) | 625 (5%) |
| # OG Representatives with contracted gene copy number | 2 (0%) | 15 (0%) | 123 (1%) |

Supplementary Table 7. Wheat gene copy number and gene retention rate in Orthologous Assemblies.

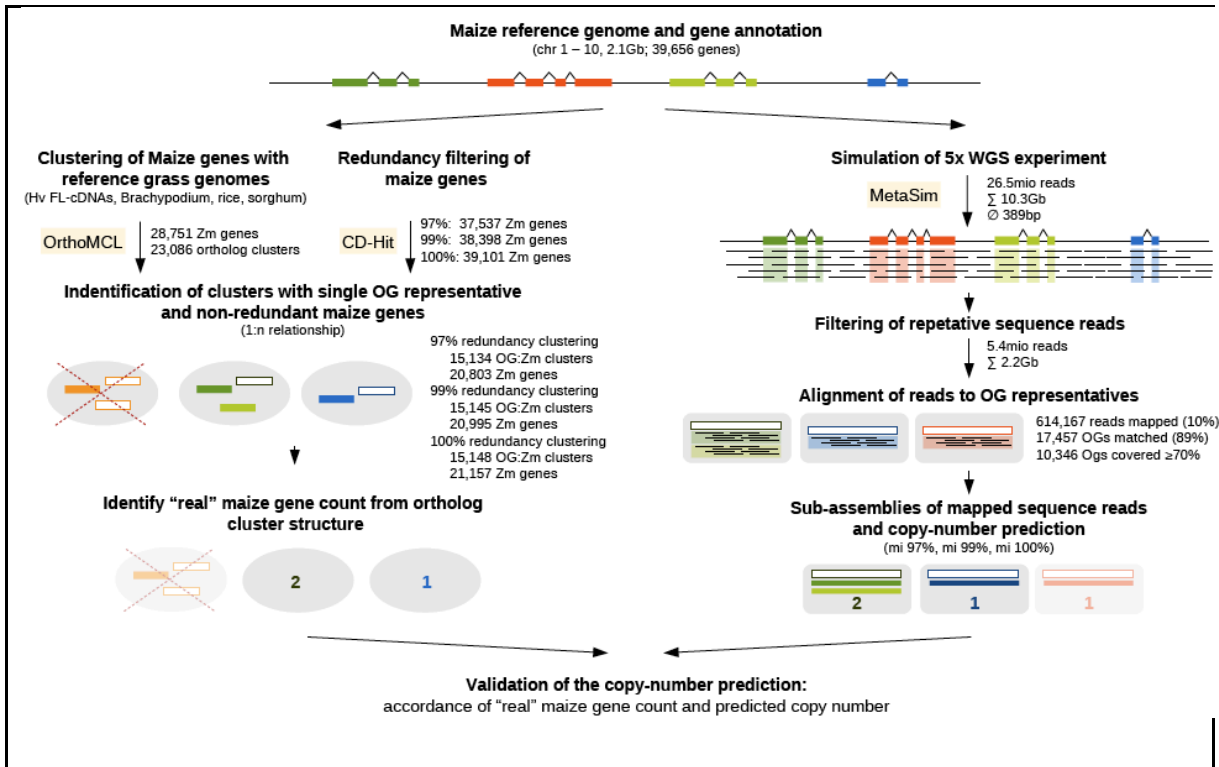
2.7 Evaluation of copy number prediction method using simulated datasets

2.7.1 Simulation using maize whole genome shotgun (WGS) data

In order to evaluate the copy number prediction, the maize reference genome sequence (ZmB73 5b.60) and its corresponding predicted gene set were used. Repeat-masked genomic sequence data (2.1Gb) and the coding sequences of 39,656 predicted genes were downloaded from <http://www.maizegdb.org>. As shown in Supplementary Figure 3 (left part), the “real” maize gene family size (gene count) was initially determined using OrthoMCL clustering of annotated maize genes with the reference grass genomes used in the orthologous grass gene set. 27,616 out of 39,656 maize genes were clustered with genes from the other grass genomes. Ortholog clusters that contained at least one maize gene and one single OG Representative from the reference genomes were identified and selected. In addition, sequence clustering of maize genes were performed at different levels (97%, 99% and 100%) using CD-Hit (-n8) (<http://www.bioinformatics.org/cd-hit/>) to

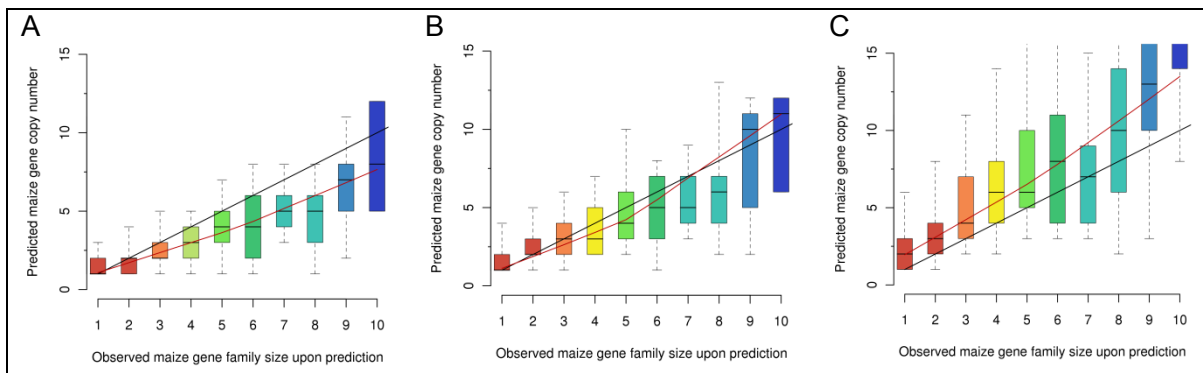
filter for redundant sequences. This strategy detected between 15,134 and 15,148 (97% to 100% redundancy clustering respectively) OG Representatives that clustered with at least one maize gene and thus were used for the evaluation. Then a 5x coverage whole genome sequencing experiment was simulated (Supplementary Figure 3, right part). 10.3Gb of sequencing reads were created using MetaSim (version 0.9.5)²² and an uniform error distribution was applied. An error rate of 0.5% and an empirical read length distribution as deduced from the wheat sequencing data used in this study was applied. The analytical workflow for computing gene-centric sub-assemblies as used for wheat analysis was carried out for the simulated maize WGS data. Mapping stringencies used were: $\geq 70\%$ sequence identity against barley, $\geq 67\%$ against *Brachypodium*, $\geq 64\%$ against rice and $\geq 68\%$ against sorghum gene representatives; at least 30 amino acids had to be included in a BLAST high-scoring segment pair (HSP). Subsequently the maize gene copy-number as detected by the simulation were computed for different assembly stringencies (mi 97%, mi 99% and mi 100%). Only OG Representatives covered by more than 70% by sub-assemblies were used for further analysis. Finally, the predicted maize gene copy number was compared to the gene family size found in the maize genome for reference. This analysis is mapped out in Supplementary Figure 3.

As illustrated in Supplementary Figure 4C, a minimum overlap identity of 100% (mi100%) clearly overestimated the gene copy number (mean ratio of polynomial median fit 1.42), whereas minimum overlap identity 97% (mi 97%) underestimated prediction (mean ratio of polynomial median fit 0.79)(Supplementary Figure 4A). Application of mi 99% showed the most balanced prediction following a one-to-one relationship between real and predicted copy number (mean ratio of polynomial median fit 0.97)(Supplementary Figure 4B). The robustness of copy number prediction decreases with an increasing reference gene count and the predicted gene family size for expanded gene families tends to be slightly underestimated. Therefore analyses were limited to wheat gene families with up to 75 copies, as beyond that genes were associated with repeat sequences, and the diploid orthologous gene family size was limited to ten genes or fewer due to reduced accuracy in estimating larger gene family sizes.



Supplementary Figure 3. Pipeline for evaluation of gene copy number prediction using simulation of a maize whole genome sequence assembly.

The methods are described in Section 2.7.1 above.

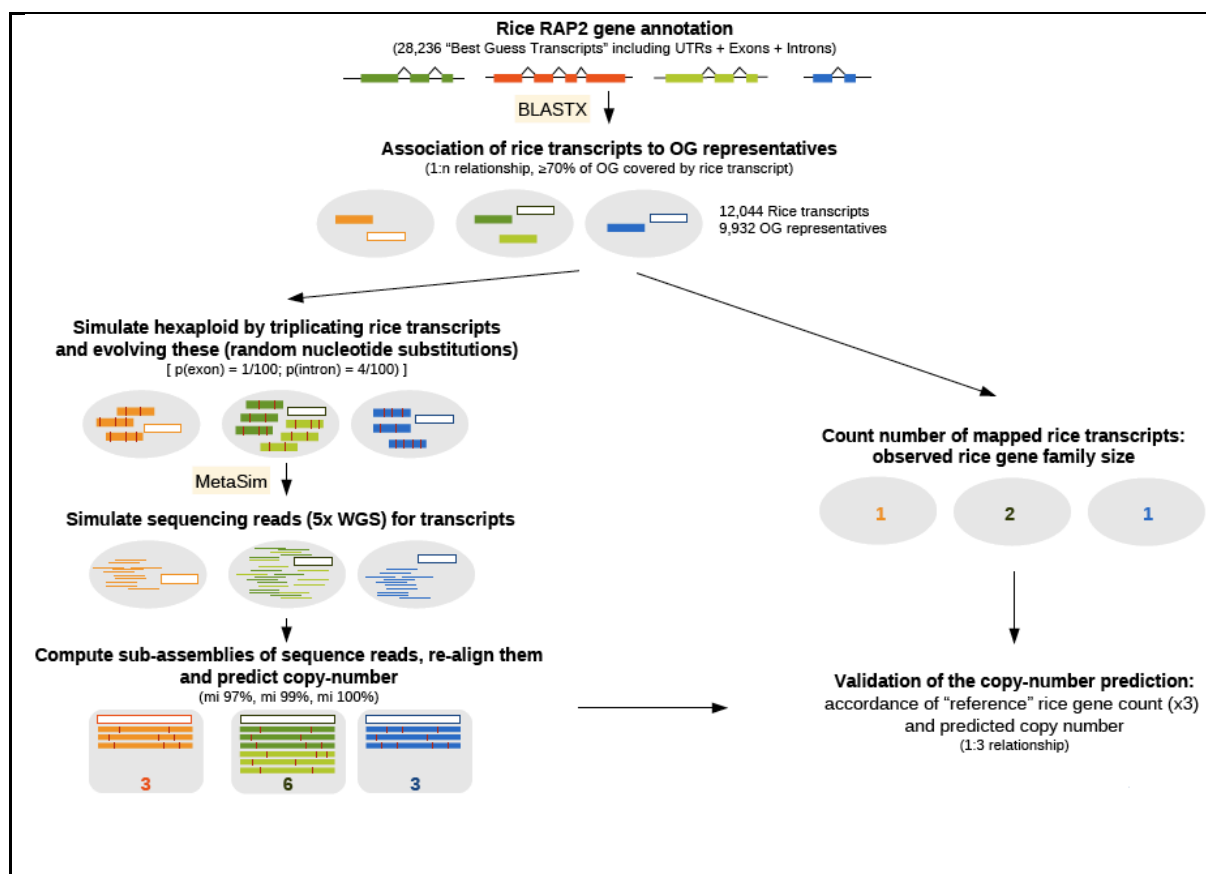


Supplementary Figure 4. Comparison of predicted gene copy numbers derived from simulations and the actual gene count of maize.

The plot shows the observed gene count against the predicted gene count based on sub-assemblies using different assembly stringencies: (A) minimum overlap identity of 97%, (B) mi 99% and (C) mi 100%.

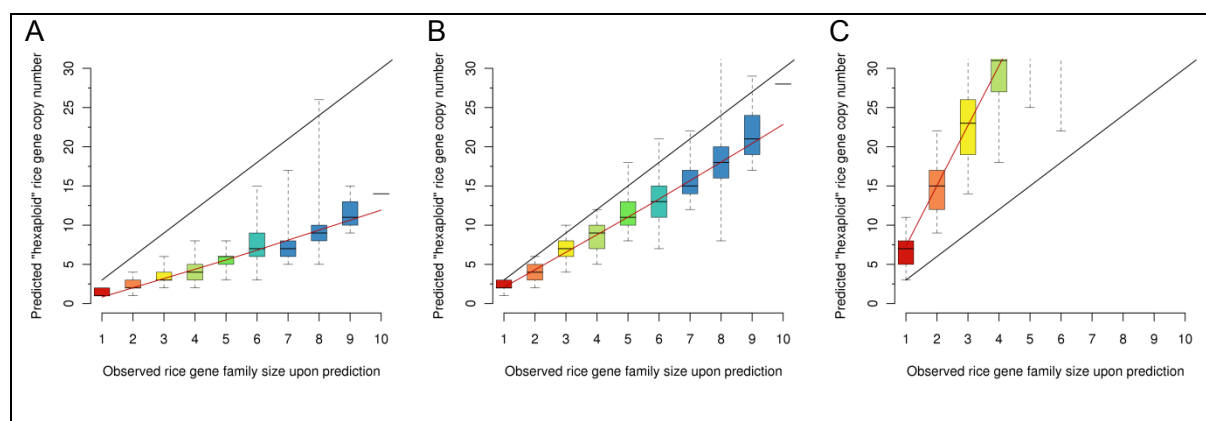
2.7.2 Validation of gene copy number prediction tested by an *in silico* generated rice hexaploid genome (`Trice`)

To evaluate the effectiveness of the OG assembly method for separating highly similar gene copies, we performed a simulation using an artificial and controlled hexaploid gene set of rice (Supplementary Figure 5). Based on the rice genome annotation (RAP2), 28,236 predicted transcripts including UTRs, exon and intron sequences were mapped against the orthologous gene representatives (OGs) using BLASTX (first-best match, $\geq 65\%$ alignment similarity against barley, $\geq 65\%$ for *Brachypodium*, $\geq 80\%$ against rice and $\geq 55\%$ against sorghum gene representatives; HSP ≥ 30 amino acids). In total, the rice transcripts tagged 15,718 OG Representatives (out of 16,160 OGs in total; 97,3 %). After further filtering for $\geq 70\%$ coverage of the OG Representative by an individual rice transcript, three copies of each rice transcript sequence were created to simulate a hexaploid gene set. Each copy was (*in silico*) evolved by random substitution of single nucleotides applying a substitution probability of 0.01 in exon and 0.04 in UTR and intron sequences. Shotgun reads were generated of the transcript sequences simulating 5x coverage using MetaSim (uniform error model accounting for a total error rate of 0.5%, uniform read length with average 389bp). Gene-centric sub-assemblies were computed using the previously described pipeline and different minimum overlap identities (mi 97%, mi 99% and mi 100%) and at least 40bp overlap. Finally, the gene copy number was predicted and compared with the observed reference gene family size that was defined as the number of rice transcripts associated to an OG Representative. This analysis is mapped out in Supplementary Figure 5 below.



Supplementary Figure 5. Pipeline for evaluation of gene copy number prediction by simulation of a hexaploid rice gene set.

The methods used are described in Section 2.7.2 above.



Supplementary Figure 6. Relationship between the observed gene family size and the predicted "hexaploid" rice gene copy number.

The plot shows the observed gene family size of rice against the predicted gene count based on sub-assemblies using different assembly stringencies: (A) minimum overlap identity of 97%, (B) mi 99% and (C) mi 100%.

The predicted gene copy number exceeds the expected 1:3 relationship when requiring perfect overlaps (mi 100%) of reads for assembly (Supplementary Figure 6C) and results in an underestimate for mi97 (Supplementary Figure 6A). Using mi99 resulted in a gene copy number estimate close to the real copy number distribution (Supplementary Figure 6B). However, frequently high similarity of the simulated reads in coding regions lead to locally collapse of homeologous genes and therefore an underestimate of the gene copy number. Nevertheless, this approach predicted the gene copy number within an interval of \pm one copy for 73% of the OG Representatives.

2.8 Transcriptome assembly

Both normalized and non-normalized cDNA were sequenced using the 454 GS FLX Titanium and long-read platforms and all read data was combined for assembly. The data set comprised 4,859,388 reads (1.6Gb) and was assembled using the Newbler gsAssembler (version 2.3) with the “-cdna” and “-large” parameters. Contigs of <100bp were discarded, and 5.6% of the remainder were identified as containing transposable elements using a BLASTn search (1e-05, -F F) against TREP and were removed. The final assembly contained 97,481 contigs of 93,430,842bp in total.

| Wheat transcriptome assembly | |
|--------------------------------|-----------------|
| # of singletons | 251,112 (5%) |
| # of assembled reads | 2,737,255 (56%) |
| # reads excluded from assembly | 1,873,424 (39%) |
| # of repeat reads | 1,588,248 |
| # of outlier reads | 73,513 |
| # too short reads | 211,663 |
| # of assembled contigs (all) | 126,555 |
| # of assembled contigs >100bp | 103,333 |
| Total sequence (bp) | 99,170,859 |
| Min / max length (bp) | 100 / 16,439 |
| Mean length (bp) | 959.72 |
| N50 / N90 (bp) | 635 / 245 |

Supplementary Table 8. Newbler assembly statistics for the wheat transcriptome assembly.

| RNA samples | CS42 Tissue and Treatment |
|------------------|--|
| Pool 1 | Immature and semi-mature Seed, Root, Flower, Young Shoots and Leaf |
| Pool 2 | 3- and 6- day salt- stressed shoots and roots and 6 day drought leaves |
| Pool 3 | Circadian- sampled seedling leaves, 0, 6, 12, 18hrs |
| Total Pooled RNA | Pool 1, Pool 2, Pool 3 |

Supplementary Table 9. Sources of RNA for cDNA analysis.

2.9 Association of wheat transcriptome data to OG Representatives

For a total of 97,481 wheat transcriptome assemblies, 54,368 (~56%) could be mapped onto at least one orthologous representative using BLASTX with the same parameters used for the mapping of the genomic sub-assemblies (see section 2.4). The remaining 43,113 transcriptome assemblies can be composed of (parts of) wheat specific genes, triticeae specific genes and non-protein-coding transcripts (nTARs, pseudogenes etc.).

The 54,368 transcriptome assemblies mapped to a total of 15,266 distinct OG Representatives, 12,615 if only first best BLAST hits are accepted. Consequently there is transcriptional support for 63% to 76% of the OG Representatives under these conditions of RNA isolation (see Supplementary Table 9). We also compared a public set of 90,786 wheat ESTs (HarvESTs) against the the OG Representatives using the same BLASTX parameters, and mapped 57,615 (63%) ESTs onto 13,103 distinct OG Representatives (10,178 if only the first-best match was considered).

2.10 Estimation of wheat gene numbers

We identified 20,051 OG Representatives with associated wheat gene assemblies. Gene copy number measurements (see section 2.5) identified 58,758 distinct copies (G_{HC}) with $\geq 70\%$ coverage of 12,481 OG Representatives ("high-coverage genes"). An additional 7,570 OG Representatives with $< 70\%$ coverage (G_{LC}) were also identified ("low-coverage genes"). We estimated the total number of wheat genes by considering average orthologous gene family size (s), gene retention rates (r) and the percentage of wheat genes not found in the OG dataset (c) as:

$$(G_{HC} + G_{LC} * s * r) / c$$

$s=1.46$; The average orthologous gene family size in sequenced diploids (*Sorghum*, rice, *Brachypodium*) (Figures 3c-e)

$c=0,92$; compensates for 8% of wheat genes not found in the OG Representatives by matching to wheat flcDNA (Supplementary Figure 2)

$r= 2,5-2,7$; the ratio of wheat gene family sizes in the hexaploid compared to diploid Triticeae (Figure 3e). The naïve expectation is a ratio of 3:1.

We therefore estimate the gene content of the wheat genome in the range of 93,900 ($r=2,5$) to 96,300 ($r=2,7$) genes.

2.11 Mapping wheat 454 reads to the *Brachypodium* genome and integration with the wheat genetic map

BLASTN (-e 1e-05) was used to compare the wheat 5X 454 read dataset to *Brachypodium distachyon* exons (version 1.2). Sequence markers that have previously been genetically mapped¹⁰ were BLAST searched against *Brachypodium* exons (v1.2). The location of the best hit was then used to anchor the genetic marker onto the *Brachypodium* genome. Where there were multiple matches, the site that conserved synteny with neighbouring markers was used.

Supplementary Figure 7. Separate maps of genetic markers on the A, B and D genomes.

These Figures are available as separate downloads.

3. Genome change in polyploid wheat

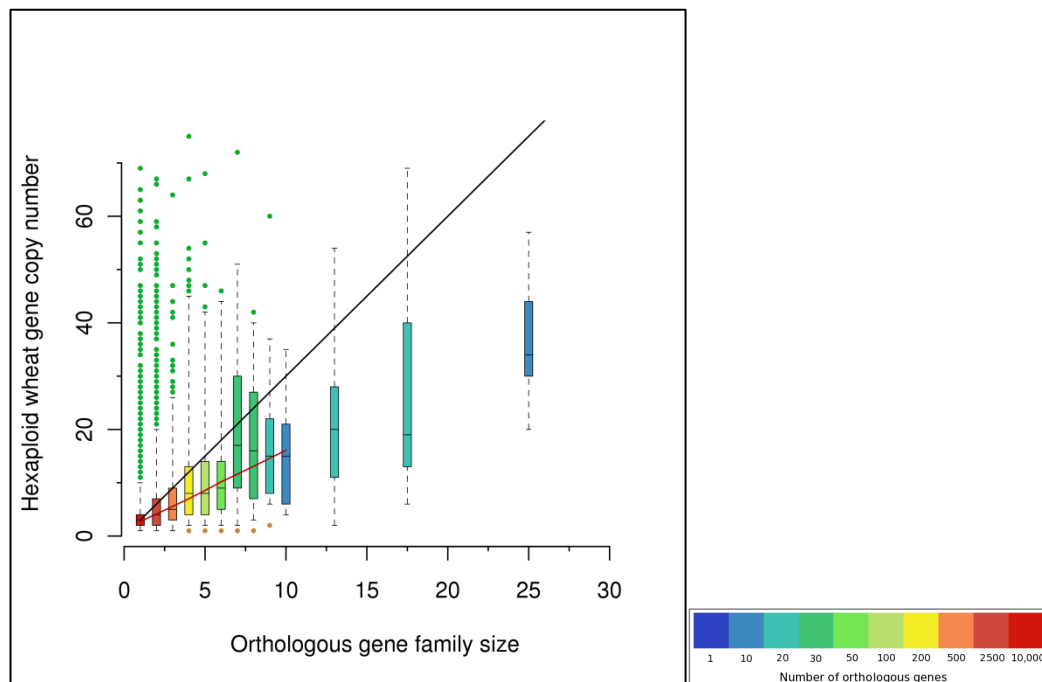
3.1 Identification of OG representatives with expanded and contracted gene copy number in wheat

For each reference gene family size we determined the frequency distribution of wheat gene copy numbers and identified the confidence interval including 90% of all OG Representatives. We identified gene family members showing significant expansion (above the 95% confidence interval) and contraction (below the 5% confidence interval). These gene family members were termed expanded and contracted families and were used for functional analysis.

3.2 Expanded and contracted gene families in *Triticeae* and wheat

We extracted PFAM domain signatures and GO terms for the gene annotations of the reference organisms *Brachypodium distachyon*, rice, barley and sorghum from SIMAP¹¹ (<http://liferay.csb.univie.ac.at/portal/web/simap>). Only GO terms from the category of molecular function were considered because of better transferability and comparability. To identify GO terms over- and under-represented in expanded and contracted OG groups

(as determined by OrthoMCL) we used the GOstats R package from Bioconductor¹² (<http://www.bioconductor.org/packages/release/bioc/html/GOstats.html>). Significant terms are reported up to a p-value of smaller than 0.05. To identify PFAM domains over- and under-represented in expanded and contracted OG groups (as determined by OrthoMCL) we used in-house software using Bonferroni correction for multiple testing. To assess GO Slim terms (molecular function category only) for lists of plain GO terms we used AgBase¹³ (http://agbase.msstate.edu/cgi-bin/tools/goslimviewer_select.pl) with the 'Plant Slim/TAIR version Aug.2011' GO Slim set.



Supplementary Figure 8. Gene retention rates for larger wheat gene families.

For each OG Representative, the gene family size was paired with the predicted wheat or *Ae. tauschii* gene copy number determined by the matching sub-assembly copy numbers. The box portion of the datasets includes 50% of the data restricted to the lower quartile and upper quartiles, and the whiskers contain in total 90% of the observed values. Outlying OG Representatives over the upper whisker boundary were defined as expanded gene families (green dots), and OG Representatives below the lower whisker boundary were defined as contracted gene families (brown dots). The colour code indicates the data density of OG Representatives. The black line shows the naïve hexaploid:diploid gene ratio of 3:1, and the red line shows a median locally-weighted polynomial regression fit through the samples up to an OG family size of 10, and represents an experimentally observed hexaploid:diploid grass ratio of 1.83:1. Wheat gene families >75 members were not analysed due to their repeat content.

Supplementary Table 10. Over- and under-represented GO terms of expanded and contracted wheat gene families.

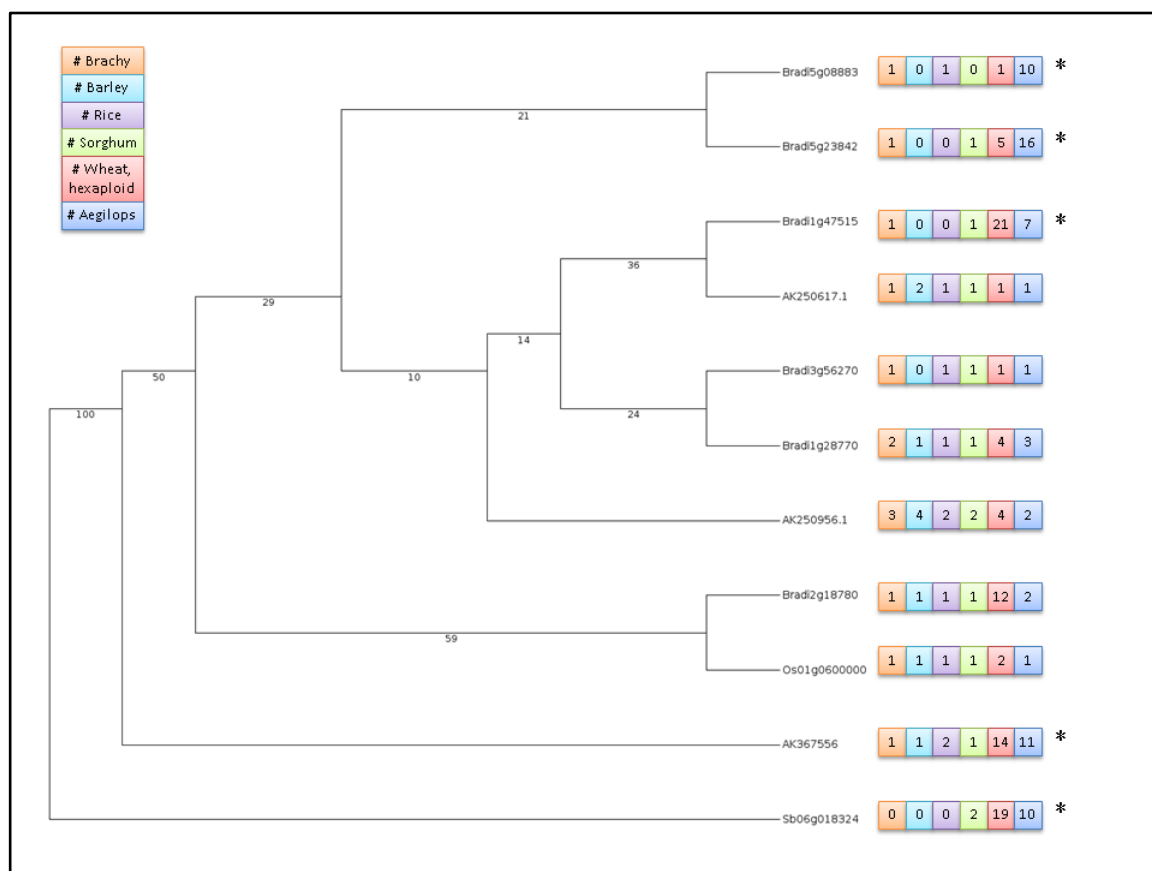
This table is available as a separate download.

Supplementary Table 11. Over- and under-represented Pfam terms of expanded and contracted wheat gene families.

This table is available as a separate download.

Supplementary Table 12. Over- and under-represented GO and Pfam terms of expanded *Ae. tauschii* gene families.

This table is available as a separate download



Supplementary Figure 9. Gene family analysis: reduction in size of the hydrogen ion transporter activity (GO:0015078) family in hexaploid wheat.

Genes were extracted from the set of OG Representatives annotated as GO:0015078 hydrogen ion transmembrane transporter activity. A total of 11 genes was identified including 5 showing significant expansion (>95% quantile) in *Ae. tauschii* compared to *Brachypodium*, Rice, Sorghum and barley (AK367556, Sb06g018324, Bradi5g23842, Bradi5g08883 and Bradi1g47515; indicated with an asterisk in the corresponding figure). We constructed a phylogenetic tree for all 11 GO:0015078 protein representatives using PROTDIST from the phylip package (bootstrapping 100 iterations). The boxes next to the

gene names indicate the copy numbers of Brachypodium, Sorghum, Rice and barley genes in the OrthoMCL group the respective gene is representing. Wheat and *Ae. tauschii* copy numbers were derived as described in Supplementary Sections 3.1 and 3.2 above. The copy numbers in wheat are derived from the hexaploid state.

4. Pseudogene Analysis

4.1 Identification of potential pseudogenes

Inspection of sub-assemblies mapped to OG Representatives identified the frequent occurrence of local “stacks” of gene fragments comprised of several distinct sub-assemblies that were not collapsed by assembly and which mapped to the same regions on their cognate OG. Local stacks were systematically identified by calculating the number of mapped sub-assemblies at each sequence position of an OG Representative using the hit count profile metric (Section 2.5). The relative mapping depth of stacks was determined by dividing each value of the hit count profile by the previously determined wheat gene copy number, and stacks were defined as regions showing at least five-fold increased mapping depth (relative mapping depth ≥ 5) over a minimum continuous stretch of 30 amino acids.

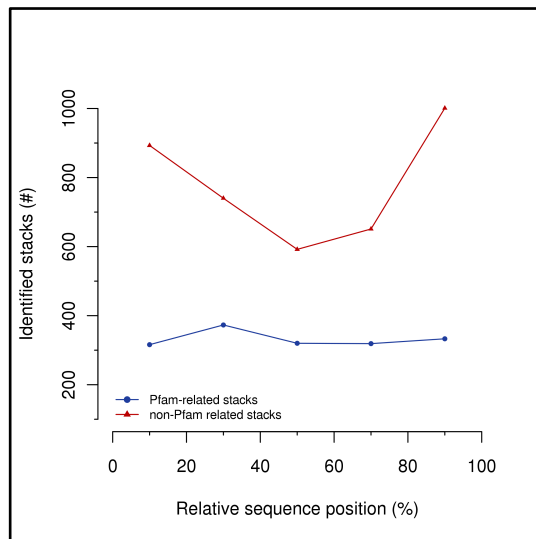
Two distinct categories of stacks were identified. Pfam-related stacks overlap in at least one sequence position with a known Pfam domain of the OG. As these stacks are associated with conserved protein domains, they may originate from genes that are not in the orthologous set. The second type of stack was not associated with any known protein domain, but are multiple fragments associated with sub-assemblies representing distinct OGs. This class of stacks was termed “pseudogenes” based on their multiple fragmentary composition, and were associated to OG Representatives at least 90% of their sequence aligned to a region identified as stacks. For the sub-assemblies, nucleotide sequence of the mapped region was extracted and translated into protein sequence. Protein alignment was re-calculated using CLUSTALW with default parameters and translated into the corresponding DNA-alignment using the corresponding CDS sequence of the OG Representative. An approximation of the maximum likelihood estimate of the synonymous substitution rate K_a (number of nonsynonymous substitutions per nonsynonymous site) and the synonymous substitution rate K_s (number of synonymous substitutions per synonymous site) was calculated using the PAML44 *yn00* package which implements the method of Yang and Nielsen¹⁴. To identify Pfam domains over- and under-represented in OG Representatives related to stacks we used the same analysis pipeline described in section 3.1. The results of this analysis are given in Supplementary Table 14.

| | Pfam-related stacks | “Pseudogene”- stacks | Σ^1 |
|---|------------------------|-------------------------|---------------|
| Analysis using all OG Representatives with sub-assemblies (ml40, mi99%) | | | |
| # of identified stacks | 2,369 | 5,543 | 7,912 |
| # of OG Representatives with stacks | 1,864 | 3,938 | 5,464 |
| Analysis using all OG Representatives with $\geq 70\%$ coverage by sub-assemblies (ml40, mi99%) | | | |
| # of OG Representatives for analysis | - | - | 12,518 |
| # of sub-assemblies for analysis | - | - | 761,470 |
| # of identified stacks | 1,661 | 3,877 | 5,538 |
| # of OG Representatives with stacks | 1,266 (10%) | 2,631 (21%) | 3,648 (29%) |
| # of sub-assemblies included $\geq 90\%$ into stacks | 69,947 (9%) | 162,930 (21%) | 232,877 (31%) |
| Mean coverage of OG Representative by stacks | 12.19% | 10.85% | 11.25% |
| Mean length of local stacks | 171bp | 163bp | 165bp |
| Mean depth of stack regions | 35.64 | 32.51 | 33.45 |
| Mean exceed of depth compared to the predicted gene copy number in stacks ² | 9.43 | 8.79 | 8.98 |
| ¹ OG representatives including PFAM and “pseudogene”-stacks were counted once | | | |
| ² depth measured as number of mapped sub-assemblies at a sequence position | | | |
| ³ mean exceed calculated as the mean ratio between depth ² and predicted gene copy number | | | |

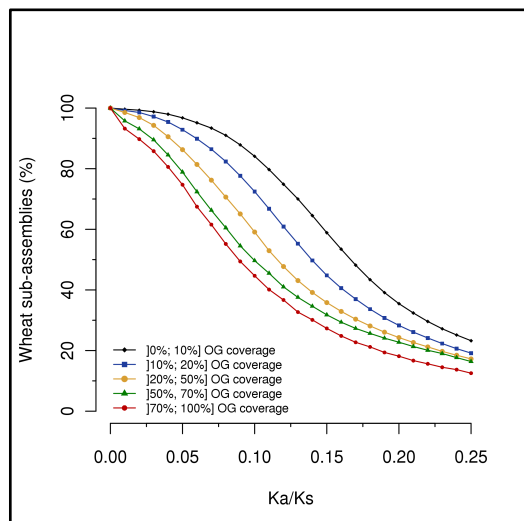
Supplementary Table 13. Analysis of “stacks” of gene fragments.**Supplementary Table 14. Pfam domains in gene fragments.**

This table is available as a separate download.

a



b



Supplementary Figure 10. Analysis of gene fragments forming “stacks”.

- The graph plots the location of stacks on OG Representatives relative to the gene structure of their OG Representative. The distribution of stacks was measured by dividing the OG Representative protein coding region into five equally sized segments (from the N terminus at 0 to the C terminus at 100) and counting the number of fragments located within each OG segment.
- The cumulative frequency distribution of sequence differences in sub-assemblies covering the coding regions of OGs to different extents is shown. Ka/Ks analyses were performed for each alignment between sub-assemblies and single-exon OG Representatives separately.

5. Determining homeologous relationships of gene assemblies

5.1 Classification of wheat sub-assemblies to the A, B or D sub-genomes

The approach taken to identify sub-assemblies of CS42 sequence as A-, B- or D-derived used the genome sequences of the D genome donor species *Ae. tauschii* and the A genome relative *Triticum monococcum*, and cDNA sequence assemblies from *Ae. speltoides*, a member of the Sitopsis section to which the putative B genome donor belongs. Varying sequence similarities of the sub-assemblies to each of these datasets would define their origin, based on the hypothesis that A- related sub-assemblies are more related to *T. monococcum* sequences, D- related sub-assemblies to *Ae. tauschii*, and B – related sub-assemblies to *Ae. speltoides*. The sequence relationships were classified by a machine- learning approach that uses the known sequences of chromosome 1A, 1B, and 1D¹⁵ to train a discriminatory kernel.

5.1.1 Defining datasets

To reduce the size of the *T.monococcum* Illumina genome sequence datasets and increase read lengths, 40% of the 101 base reads were sub-sampled and assembled with SOAPdenovo (<http://soap.genomics.org.cn/soapdenovo.html>) by using k-mer sizes ranging between 45 and 61bp. Returned contigs with less than 100bp sequence length were removed to exclude assembly artifacts. For each k-mer size the N50 was calculated to assess the assembly quality. A k-mer size of 61bp showed a maximum N50 of 204bp and was taken as final assembly for further analysis. The *Ae. tauschii* genome set comprised the 3x genome coverage with 454 reads (J. Dvorak, unpublished data), and a Trinity assembly of *Ae. speltoides* cDNA (Trick and Bancroft, unpublished data). The wheat whole-genome sub-assemblies (assembled at 99% identity) associated with one representative reference gene model from the ortholome set that had hits to all three datasets above (692,631 (73%) sub-assemblies) were used to be classified. To train the machine learning algorithm, sequences from flow sorted chromosomes of wheat 1A, 1B and 1D were used¹⁵.

| <i>T.monococcum</i> (A- related) | |
|---|-----------------------------------|
| Whole genome shotgun sequence data | WGS Illumina (Paired-End) |
| total sequence [Gbp] | 400.48 |
| Read length | 36bp / 101bp |
| Insert Sizes | 220bp / 280bp / 2,000bp / 5,000bp |
| Assembly | 65,851,441 |
| used input data (40% of 101bp reads) | 118.97 Gbp |
| Library 1 (sequence, read length, insert size) | 71.49 Gbp, 101bp, 220bp |
| Library 2 (sequence, read length, insert size) | 71.48 Gbp, 101bp, 280bp |
| SOAPdenovo assembly, kmer size 62, (min contig length 100bp) | 460 / 260 |
| total sequence [Mbp] | 3,687.8 |
| minimum / maximum length [bp] | 100 / 20,526 |
| mean length [bp] | 208.7 |
| N50 [bp] | 204 |
| | |
| <i>Ae.speltoides</i> (Sitopsis section) | |
| Type | cDNA Assembly (Trinity) |
| total sequence [Mbp] | 151.4 |
| minimum / maximum length [bp] | 200 / 17,190 |
| mean length [bp] | 895.83 |
| N50 [bp] | 1,645 |
| | |
| <i>Ae.tauschii</i> (D- genome donor) | |
| Type | WGS 454 reads |
| total sequence [Mbp] | 12,800.1 |
| minimum / maximum length [bp] | 40 / 2,044 |
| mean length [bp] | 377.23 |
| N50 [bp] | 447 |

Supplementary Table 15. Reference data sets for ABD classification.**5.1.2 Construction of the training set**

To define a suitable machine learning training set we made use of known wheat group 1 chromosome sequences which were separated into their sub-genomes (A, B and D) using flow sorting¹⁸. First we extracted all wheat sub-assemblies associated with wheat group 1

chromosomes. These were identified as those within syntenic blocks defined by alignment of barley chromosome 1 to *Brachypodium*, rice and sorghum¹⁶.

A total of 1,607 representative genes (OGs) with 18,684 associated sub-assemblies was selected as a training-set after additional filtering (criteria: subassembly hits to all 1A, 1B and 1D; one hit > 95% identity; ≥ 0.01 difference in sequence similarity ratio; sub-assembly hits to all *T. monococcum*, *Ae. speltooides* and *Ae. tauschii*). Each of these filtered chromosome group 1 sub-assemblies was labeled A, B or D depending on its best hit to chromosome 1A, chromosome 1B and chromosome 1D reads. Out of the 18,684 sub-assemblies, 6,185 were assigned to (1)A, 6,326 to (1)B and 6,175 to (1)D.

The resulting chr1 classification had the following exemplary characteristics:

```
assemb_id, sim_to_1A, sim_to_1B, sim_to_1D, class_assigned
subassembly_XYZ, 92.76, 97.60, 94.07, B
```

Each sub-genome classification was then complemented with the sub-assemblies' similarities to *T. monococcum*, *Ae. speltooides* and *Ae. Tauschii* sequences to create a training set compatible to all non-chr1 related sequences:

```
assemb_id, sim_to_Mono, sim_to_Spelt, sim_to_Aegil, class_assigned
subassembly_XYZ, 91.76, 95.60, 92.07, B
```

Several machine learning algorithms from the WEKA package (<http://www.cs.waikato.ac.nz/ml/weka/>)¹⁷ were applied to this training set and the results were evaluated by stratified cross-validation. The results from five machine learning algorithms are shown in Supplementary Table 16.

| Machine Learning method | Correctly classified total | Correctly classified % |
|---------------------------------------|----------------------------|------------------------|
| Decision Tree | 10,846 | 58.0497% |
| Logistic Regression | 10,494 | 56.1657% |
| Naïve Bayes | 10,752 | 57.5466% |
| Support Vector (SMO) | 10,388 | 55.5984% |
| Support Vector (libSVM) ¹⁸ | 11,076 | 59.2807% |

Supplementary Table 16. Machine learning algorithms applied to training set.

The Receiver- Operator Characteristics of the two Support Vector classifiers are shown below:

1. Support Vector Machine (SMO), default options.

```
=== Detailed Accuracy By Class ===
```

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.591 | 0.242 | 0.547 | 0.591 | 0.568 | 0.681 | A |
| | 0.421 | 0.177 | 0.549 | 0.421 | 0.477 | 0.629 | B |
| | 0.659 | 0.246 | 0.569 | 0.659 | 0.611 | 0.71 | D |
| Weighted Avg. | 0.556 | 0.221 | 0.555 | 0.556 | 0.551 | 0.673 | |

2. Support Vector Machine (libSVM), options: -S 0 -K 2 -D 3 -G 0.05 -R 0.0 -N 0.5 -M 40.0 -C 2.0 -E 0.0010 -P 0.1 -B

```
=== Stratified cross-validation ===
```

| | | |
|----------------------------------|-----------------|----------------------|
| Correctly Classified Instances | 11076 | 59.2807 % |
| Incorrectly Classified Instances | 7608 | 40.7193 % |
| Kappa statistic | 0.3887 | |
| K&B Relative Info Score | 486166.4408 % | |
| K&B Information Score | 7705.1266 bits | 0.4124 bits/instance |
| Class complexity order 0 | 29611.8036 bits | 1.5849 bits/instance |
| Class complexity scheme | 24958.5941 bits | 1.3358 bits/instance |
| Complexity improvement (Sf) | 4653.2095 bits | 0.249 bits/instance |
| Mean absolute error | 0.3678 | |
| Root mean squared error | 0.4257 | |
| Relative absolute error | 82.7706 % | |
| Root relative squared error | 90.3064 % | |
| Total Number of Instances | 18684 | |

```
=== Detailed Accuracy By Class ===
```

| | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|---------|---------|-----------|--------|-----------|----------|-------|
| | 0.531 | 0.16 | 0.621 | 0.531 | 0.573 | 0.735 | A |
| | 0.642 | 0.28 | 0.54 | 0.642 | 0.586 | 0.73 | B |
| | 0.604 | 0.171 | 0.635 | 0.604 | 0.619 | 0.77 | D |
| Weighted Avg. | 0.593 | 0.204 | 0.598 | 0.593 | 0.593 | 0.745 | |

```
=== Confusion Matrix ===
```

| a | b | c | <-- classified as |
|------|------|------|-------------------|
| 3286 | 1870 | 1029 | a = A |
| 1152 | 4061 | 1112 | b = B |
| 850 | 1595 | 3729 | c = D |

The TP (True Positive) rate and Precision differ between the two algorithms, accounting for the different numbers of classified sub-assemblies. The Support Vector Machine (libSVM) showed the best compromise between precision and recall, and different parameter settings and combinations were assessed prior using it.

5.1.3 Classification of the whole-genome sub-assemblies

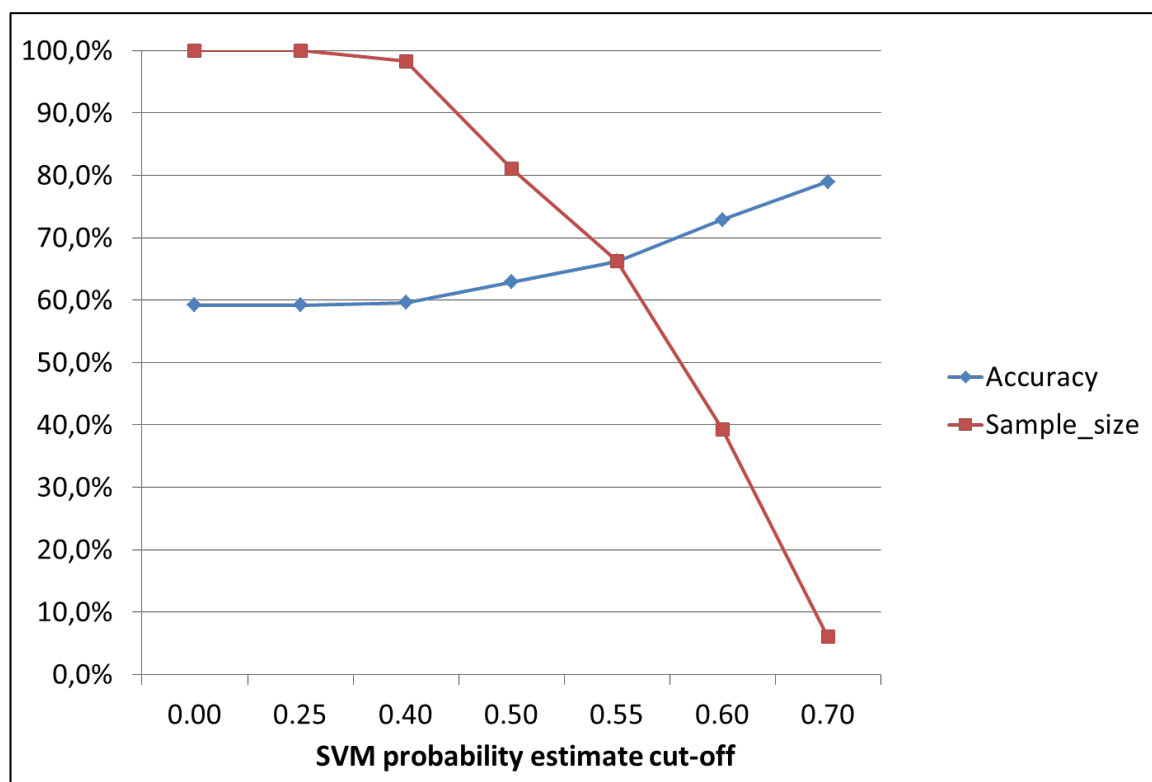
The trained libSVM classifier was applied to the complete set of whole-genome sub-assemblies with hits to all three A-, B- and D- related sequence sets (692,631 sub-assemblies). The distribution of predictions for the three classes was:

predicted_A: 176,730 (26%)

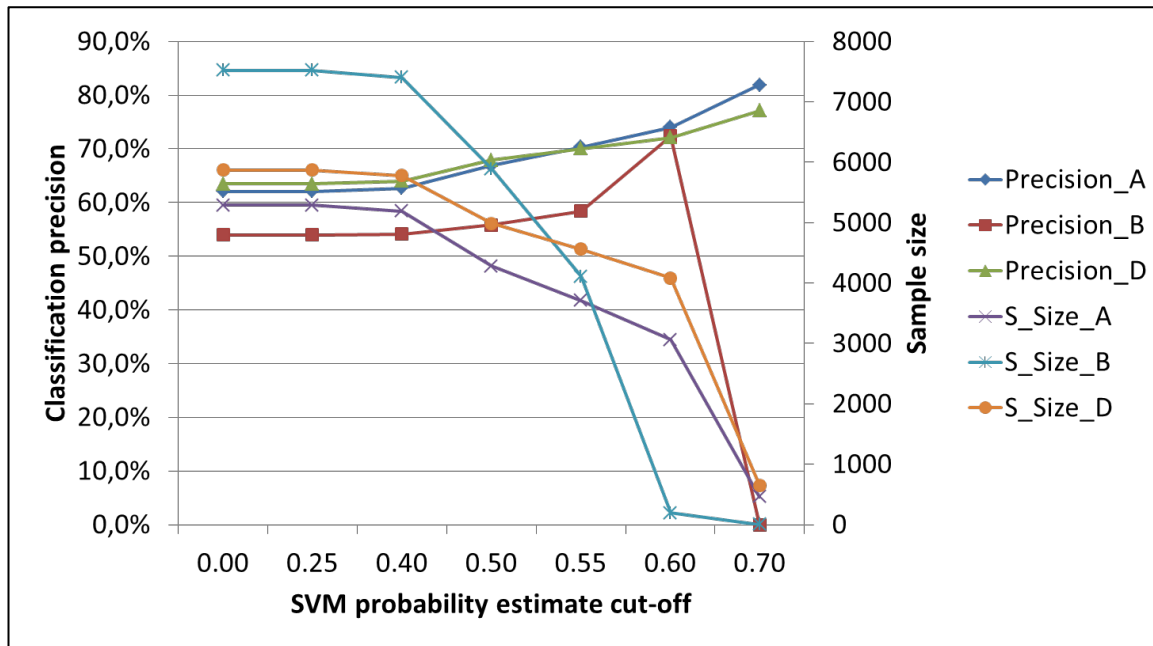
predicted_B: 306,735 (44%)

predicted_D: 209,166 (30%)

The libSVM probability estimates were used to filter for high confidence predictions while maximizing sample sizes. To determine a reasonable trade-off between accuracy and sample size, we plotted prediction accuracy versus sample size for varying probability cut-offs for the training set (Supplementary Figures 11 and 12). A probability estimate cut-off of 0.55 was selected based on these data.



Supplementary Figure 11. Accuracy versus sample size for the trained libSVM classifier.

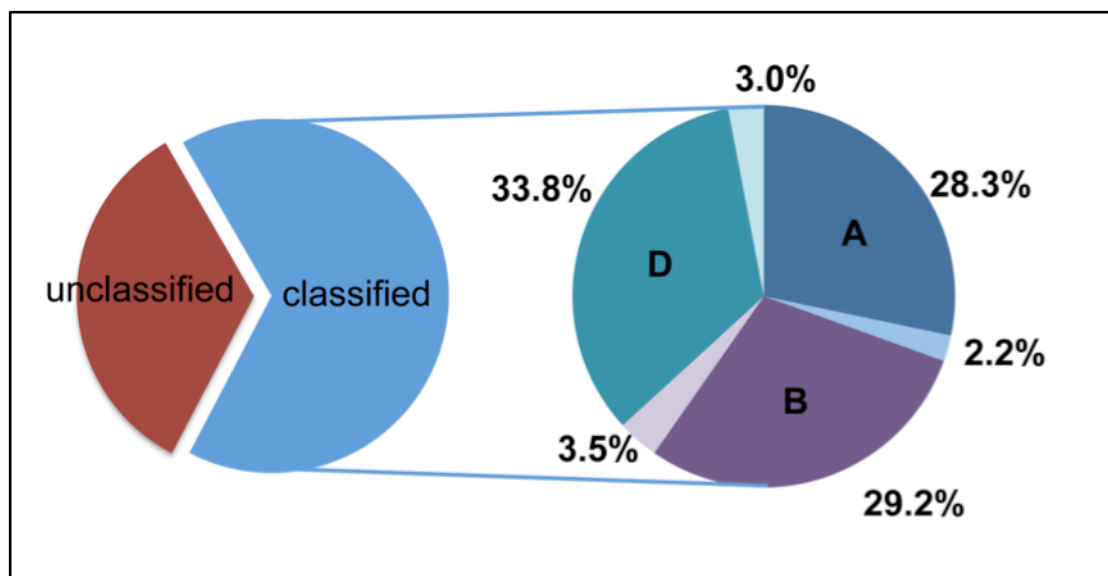


Supplementary Figure 12. A/B/D precision versus sample size for the libSVM classifier.

This set of optimized parameters was used to classify a set of 510,044 sub-assemblies that represented genes and not gene fragments and potential pseudogenes (see section 4.1). In the final analysis, classified sub-assemblies with nonsense mutations were plotted separately from the classified sub-assemblies without stop-codons to estimate the proportion of non-functional but recognizable genes in each genome. Supplementary Table 17 shows the results of the classification, unclassified sub-assemblies did not meet the minimum probability estimate cut-off of 0.55. GO terms for each classified wheat sub-assembly were determined based on their corresponding OG and the total distribution of GO Slim terms (molecular function category only) was plotted separately for A,B and D-assigned sub-assemblies in Supplementary Figure 13. The distribution of GO Slim terms (molecular function category only) for A,B and D-classified sub-assemblies with and without nonsense mutations is plotted in Supplementary Figure 14. For each GO Slim category the ratio between the percentage sub-assemblies with and without stop codons was plotted to demonstrate the potential for conservation of intact genes in different functional categories in each genome.

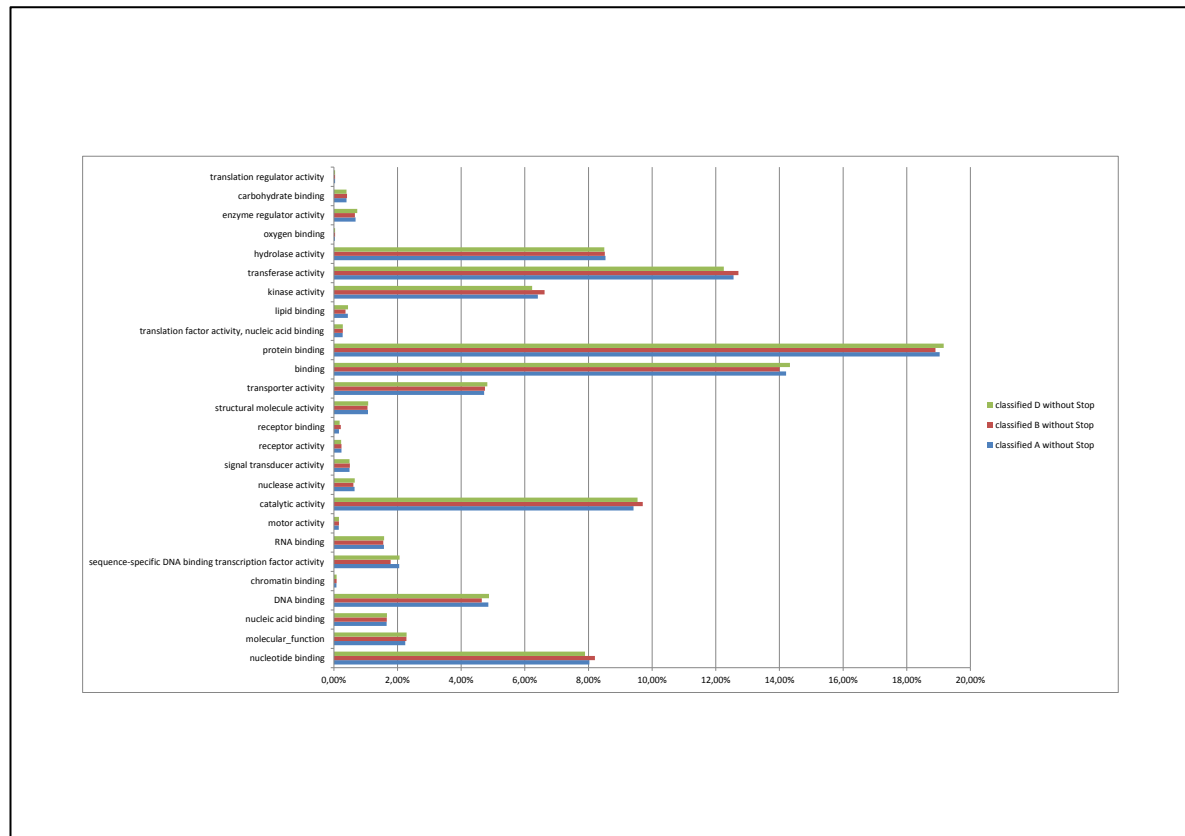
| | Stop Codons | Intact ORF | Total |
|-------------------------|-----------------|------------------|---------|
| Used for classification | 47,241 (9%) | 462,803 (91%) | 510,044 |
| Classified as A | 7,502 (7.33%) | 94,949 (92.67%) | 102,351 |
| Classified as D | 9,908 (8.06%) | 113,065 (91.94%) | 122,973 |
| Classified as B | 11,370 (10.72%) | 97,923 (89.28%) | 109,453 |
| Not classified | 18,101 (10%) | 157,166 (90%) | 175,267 |

Supplementary Table 17. Support Vector Machine classification results on wheat sub-assemblies.

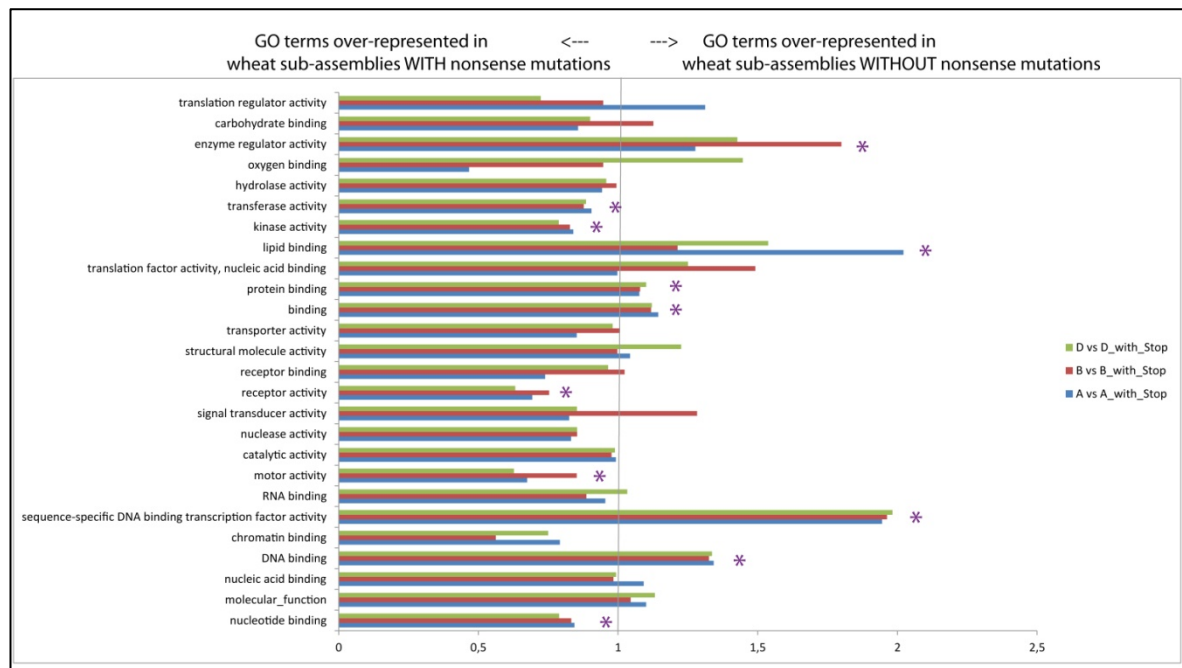


Supplementary Figure 13. Assignment of genes to A, B and D genomes.

The proportion of sub-assemblies classified by SVM is shown. The right chart shows the sub-assemblies in the A, B and D genomes. The smaller segments show sub-assemblies with stop codons.



Supplementary Figure 14. Distribution of GO Slim molecular function categories of wheat sub-assemblies assigned to the A, B, D genomes.



Supplementary Figure 15. Distribution of GO Slim molecular function categories with and without nonsense mutations. For each GO Slim category the proportion of ‘intact’ sub-assemblies and ‘disrupted’ sub-assemblies in each genome is shown. The percentage of sub-assemblies in each GO Slim category is shown in Supplementary Figure 13. Purple stars beside a category indicate statistically significant (< 0.05) two-sided p-values determined by Fisher’s Exact Test for the combined observation (A, B and D combined with and without nonsense mutation).

5.2 Identification of homeologous SNPs in Chinese spring and assignment to the A, B or D sub-genomes

5.2.1 Creation of CS reference sequences

A set of reference sequences was created by re-assembling the OA contigs and singletons for each individual orthologous assembly using CAP3¹⁹ with permissive parameters (-o 16 -p 66 -s 251 -g 1 -z 1). This collapsed homeologous sequences within each gene assembly, reducing redundancy and allowed more reads to map uniquely. Testing with a small subset of the assemblies and read data suggested this method increased the number of uniquely mapping reads from ~40% to >80% and improved reference coverage from <10% to 40%, thus providing more positions of high coverage for more confident SNP calling. The re-assembly gave a new reference comprising 313,556 sequences of 196Mb in total with a mean length of 624bp.

5.2.2 Mapping SOLiD and 454 reads to the CS reference

BWA²⁰ was used to map 6 slides (3 full runs) of CS SOLiD reads to the reference (bwa aln -n 10 -o 0 -c). Non-mapping, non-uniquely mapping and possible PCR duplicate reads were filtered out of the SAM files. An additional 18 runs of SOLiD read data from the sequencing of 3 other hexaploid wheat varieties (Avalon, Savannah and Rialto) was used to increase the coverage over the CS reference and improve the SNP identification. The mpileup program in the SAMtools package²¹ was used to determine polymorphic positions from the CS SOLiD mapping and, where the same variant appeared in the other 3 varieties, the counts for reference and novel alleles were added to CS. This ensured only homeologous and not varietal SNP positions were included. The CS 454 reads from the 5X dataset were also mapped as 50bp fragments using Bowtie²² (-f -S -a --best -v 3), and the bases at polymorphic positions were combined.

5.2.3 Identifying homeologous SNPs in CS

SNPs were called using a custom pipeline to identify them in a hexaploid, as most programs are designed for diploid genomes. Our pipeline uses the SAMtools mpileup output file of base calls and coverage (using the command for precise depth of coverage without any cut-offs imposed “-BQ0 -d10000000”). This output was first parsed to remove poor quality bases, with a score of < Phred10. Variants were initially called using non-strict parameters: at least 2 reads matching reference base; at least 2 reads matching the alternative base; the coverage of each alternative base must be $\geq 10\%$ of the total for that position. This last parameter discounts the non-reference bases that occur due to errors in sequencing but also to allow for under-representation of an alternative allele in the library. The coverage depth was restricted to between 23X and 83X (q50 and q75 of coverage) and all SNPs within 5bp of another were rejected as possible INDELs rather than single base changes. A total of 987,909 SNPs were identified in CS.

5.2.4 Assignment of CS SNPs to the A, B or D subgenomes

In this analysis, Illumina sequence reads from the A genome relative *Triticum monococcum* (shortened to “Tm”) (see Supplementary table 2) were used to represent the A genome of CS and SOLiD reads from the D genome donor *Ae. tauschii* (shortened to “At”) (Suppl. Table 2) were used to represent the D genome of CS. Both read sets were separately mapped to the CS references and results filtered as described previously. SAMtools mpileup produced a base-calls file for the Tm and At mappings and bases with quality scores < Phred10 were removed. A set of 619,022 homeologous SNPs was used in the comparison between the CS42, Tm and At genome sequences. This was the subset of the total SNPs found in CS42 where there was also coverage at the same position in both the Tm and At genome mappings. Positions with >2 different bases present in CS

(9,693 positions) were discounted, as these could appear due to mapping of reads from paralogous genes.

A custom script determined whether each CS variant was present in Tm or At or both. A CS SNP present in Tm but not At was assigned to the A genome and a SNP present in At but not Tm was assigned to the D genome. If neither genome contained the alternative base it was assigned to the putative B genome. When a CS variant was found from either, or both, of the Tm or At mappings we required that there should be a coverage depth of ≥ 5 in the Tm and At mappings otherwise the SNP was filtered out. All filtering parameters are summarized in Supplementary Table 18. SNPs are shown in Figure 2 and Supplementary Figure 7.

| Dataset | Number of SNPs |
|---|----------------|
| Chinese Spring SNPs called (coverage $>23X$, $<83X$, none within 5bp) | 987,909 |
| Position is also covered by Tm & At reads | 639,141 |
| Only 1 alternative base was found in the CS SOLiD reads | 629,448 |
| Only reference and alternative bases in Tm & At (no new bases) | 619,022 |
| Only homozygote positions in Tm & At | 462,018 |
| $\geq 5X$ mapping coverage in Tm & At | 276,184 |
| $\geq 90\%$ reads agree at Tm & At posn (allow error in 1/10 reads) | 269,677 |
| A genome | 38,703 |
| D genome | 63,890 |
| B genome inferred (SNP is in CS reference – more reliable) | 29,959 |
| B genome inferred (SNP is in SOLiD reads – less reliable) | 137,125 |

Supplementary Table 18. Filtering steps used to compare SNPs found in CS with those in *T. monococcum* (Tm) and *Ae. tauschii* (At).

The SNPs used as the final high quality set (total of 132,552) are in bold text. These are displayed on Figure 2.

5.2.5 Validation of SNP assignments

SNP cross-validation used a small set of Illumina Nimblegen array captured data for *T. urartu* (AA), two *T. dicoccoides* (AABB) lines, Paragon (AABBDD), *Ae. speltoides* (B genome-like), and two *Ae. tauschii* (DD) lines. Illumina data was mapped to the sub-assemblies. 4,408 A, B, D genome SNPs were called with complete consistency to SNP calls described above. SNP calls in the D genome were completely concordant between the two methods (858/858). For the A genome, the agreement was lower at 81% (623/766), probably due to the different A genomes used.

5.3 Assessment of Machine Learning classification of sub-assemblies to genomes using SNP identification in the CS42, the A and D genomes

Of the OA sub-assemblies that had been classified as belonging to either the A, B or D genomes (334,777 sequences), 73,547 were used in the reference sequences for mapping the CS (with Avalon, Rialto, Savannah), *T. monococcum* and *A. tauschii* data. The remainder of the OA assemblies had been re-assembled using CAP3 to reduce redundancy in the reference. 12,864 SNPs were located on 6,417 of the OA-based reference sequences and the proportions in agreement between the SVM and SNP analyses for each genome are displayed in Supplementary Table 19.

| | | Assignment from SNP analysis | | | |
|---------------------|---|------------------------------|-------------|--------------|------|
| Assignment from SVM | A | 1501 (71.9%) | 346 (16.6%) | 238 (11.4%) | 2085 |
| | B | 238 (22.6%) | 452 (42.8%) | 365 (34.6%) | 1055 |
| | D | 142 (4.3%) | 333 (10.1%) | 2802 (85.3%) | 3277 |

Supplementary Table 19. Validation of SVM sub-genome assignments using SNP assignments.

References

- 1 Sears, E. R. *Nullisomic-tetrasomic combinations in hexaploid wheat.*, 22-45 (Oliver and Boyd, 1966).
- 2 Paux, E. *et al.* Physical mapping in large genomes: accelerating anchoring of BAC contigs to genetic maps through in silico analysis. *Functional & Integrative Genomics* **8**, 29-32 (2008).
- 3 Peterson, D. G., Tompkins, J. P., Frisch, D. A., Wing, R. A. & Paterson, A. H. *Construction of Plant Bacterial Artificial Chromosome (BAC) Libraries: An Illustrated Guide.* Second edition (2002).
- 4 Onate-Sanchez, L. & Vicente-Carbajosa, J. DNA-free RNA isolation protocols for *Arabidopsis thaliana*, including seeds and siliques. *BMC Res Notes* **1**, 93, (2008).
- 5 Zhulidov, P. A. *et al.* Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Research* **32**, e37, (2004).
- 6 Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research* **35**, 7188-7196, (2007).
- 7 Li, L., Stoeckert, C. J., Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**, 2178-2189 (2003).
- 8 van Dongen, S. A cluster algorithm for graphs. (Amsterdam, 2000).
- 9 Mochida, K., Yoshida, T., Sakurai, T., Ogihara, Y. & Shinozaki, K. TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. *Plant Physiology* **150**, 1135-1146, (2009).
- 10 Allen, A. M. *et al.* Transcript-specific, single-nucleotide polymorphism discovery and linkage analysis in hexaploid bread wheat (*Triticum aestivum* L.). *Plant biotechnology journal* **9**, 1086-1099, (2011).
- 11 Rattei, T. *et al.* SIMAP-a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Research* **38**, D223-226, (2010).

- 12 Beissbarth, T. & Speed, T. P. Gostat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics (Oxford, England)* **20**, 1464-1465, (2004).
- 13 McCarthy, F. M. *et al.* AgBase: a unified resource for functional analysis in agriculture. *Nucleic Acids Research* **35**, D599-603, (2007).
- 14 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
- 15 Wicker, T. *et al.* Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *The Plant Cell* **23**, 1706-1718, (2011).
- 16 Mayer, K. F. *et al.* Unlocking the barley genome by chromosomal and comparative genomics. *The Plant Cell* **23**, 1249-1263, (2011).
- 17 Frank, E., Hall, M., Trigg, L., Holmes, G. & Witten, I. H. Data mining in bioinformatics using Weka. *Bioinformatics (Oxford, England)* **20**, 2479-2481, (2004).
- 18 Chang, C.C. & Lin, C.-J. LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **27**, 1-27 (2001).
- 19 Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Research* **9**, 868-877 (1999).
- 20 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754-1760, (2009).
- 21 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078-2079, (2009).
- 22 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25, (2009).